Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 97 (2006) 2221–2241
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Application of modified information criterion to multiple change point problems

Jianmin Pan[a,*], Jiahua Chen[b]

[a]*Department of Biostatistics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA*
[b]*Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ont., Canada N2L 3G1*

Received 27 January 2005
Available online 24 July 2006

Abstract

The modified information criterion (MIC) is applied to detect multiple change points in a sequence of independent random variables. We find that the method is consistent in selecting the correct model, and the resulting test statistic has a simple limiting distribution. We show that the estimators for locations of change points achieve the best convergence rate, and their limiting distribution can be expressed as a function of a random walk. A simulation is conducted to demonstrate the usefulness of this method by comparing the powers between the MIC and the Schwarz information criterion.

AMS 1991 subject classification: Primary 62H15; Secondary 62H10

Keywords: Consistency; Convergence rate; Limiting distribution; Multiple change points; Model complexity; Regular parametric model

1. Introduction

Information criteria are commonly used for selecting competing statistical models. Out of several competing statistical models, we do not always choose the one with the best fit to the data. Such models may simply interpolate the data and have little interpretable value. Model complexity is an important factor in information criteria for model selection, see [1,18]. The model complexity in existing criteria is often measured in terms of the dimensionality of the parameter space. Although this notion is well found in regular parametric models, it lacks some desirable properties when applied to irregular statistical models. Chen et al. [4] refined the notion

* Corresponding author. Fax: +1 901 495 4585.
E-mail addresses: jianmin.pan@stjude.org (J. Pan), jhchen@uwaterloo.ca (J. Chen).

0047-259X/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2006.05.009

of model complexity in the context of single change point problems, and modified the existing information criteria. They showed that the modified information criterion (MIC) is consistent in selecting the correct model and has simple limiting behavior. We generalize the MIC in [4] so that it can be applied to multiple change point models in this paper.

Consider the problem of making inference on whether a process has undergone some changes. In the context of model selection, we want to choose between a model with a single set of parameters, or a model with two or more sets of parameters plus the locations of changes.

Compared to usual model selection problems, the change point problem contains some special parameters: the locations of the changes. When some of them approach the beginning or the end of the process or cluster somewhere in the process, one or more sets of the parameter become completely redundant, and the model is unnecessarily complex. Hence, the model complexity should be considered as a function of both the locations of the change points and the dimensionality of the parameter space.

The change point problem has been extensively discussed in the literature in recent years. The study of the change point problem dates back to Page [16,17] which tested the existence of single change point, and Chernoff and Zacks [5] which was motivated by consideration of a "tracking" problem. Multiple change point problems also have been considered by many authors including Yao [24], Yao and Au [25], Fu and Curnow [8], Bai and Perron [2], Lee [13], Siegmund [21] and Ninomiya [15]. The problem was also discussed in a Bayesian framework, see [5,23,3,14]. The discussion of change point problem for dependent observations can be found in [12,11]. The present study deviates from other studies by refining the traditional measure of the model complexity.

Suppose we have a sequence of independent observations $X_1, \dots, X_n$. It is assumed that there exist up to R integers $\tau_1, \dots, \tau_R$, where $0 = \tau_0 < \tau_1 < \cdots < \tau_R < \tau_{R+1} = n$, such that X_i has density function $f(x, \theta_r)$ when $\tau_{r-1} < i \leqslant \tau_r$ $(r = 1, \dots, R+1)$ which belong to the same parametric distribution family $\{f(x, \theta); \theta \in \Theta\}$ with $\Theta \subset \mathcal{R}^d$.

The problem is then to test whether the R changes have indeed occurred and to estimate the locations of the R changes if they exist. For this purpose, we adopt the MIC proposed by Chen et al. [4]. It is believed that when $\tau_1, \dots, \tau_R$ are distributed evenly between 1 and n, the model is least complex and all parameters $\theta_1, \dots, \theta_{R+1}$ are effective. When one or more change points are near 1 or n, or cluster, some of parameters $\theta_1, \dots, \theta_{R+1}$ become redundant. Hence, some $\tau_1, \dots, \tau_R$ are increasingly undesirable parameters and the model is considered as the most complex in this case. To simplify notation, let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_{R+1})$ and $\boldsymbol{\tau} = (\tau_1, \dots, \tau_R)$ be the parameter vector and the location vector of change points, and use triplet $(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$ to identify the number of copies of $\boldsymbol{\theta}$'s in the model under consideration. We denote the log-likelihood function as

$$l_n(\boldsymbol{\theta}, \boldsymbol{\tau}, R) = \sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r} \log f(X_i, \theta_r).$$

The MIC for the multiple change points is defined as

$$MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R) = -2l_n(\boldsymbol{\theta}, \boldsymbol{\tau}, R) + (R+1)d \log n + C \sum_{r=1}^{R+1} \left(\frac{\tau_r - \tau_{r-1}}{n} - \frac{1}{R+1} \right)^2 \log n,$$

where $C > 0$ is a constant. Note that this criterion favors change point models with change points spreading out uniformly. This notion in single change point case is shared by many researchers.

The method in [10] scales down the statistic when the suspected change point is near 1 or n. The U-statistic in [9] is scaled down by multiplying a factor $\tau(n-\tau)$ when τ is the location of the change. From a different angle, the modification can also be used to reflect some belief on uniformity in the change points. Thus, our method also has a link to Lee [14] who showed that under uniform prior, the locations of the change points are estimated with a convergence rate of $O_p(\log n)$.

When there is no change point, we define

$$MIC(\theta, n, 0) = -2l_n(\theta, n, 0) + d \log n.$$

Let

$$MIC(\boldsymbol{\tau}, R) = \inf_{\boldsymbol{\theta}} MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R).$$

We select the model with corresponding $\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, \hat{R}$ minimizing $MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$. That is

$$MIC(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, \hat{R}) = \inf MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R) \tag{1}$$

among all choices of $(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$. When R is large, the evaluation of this criterion is a non-trivial task.

We assume the number of change points R as fixed in this paper. Further research is needed to investigate the consistency of $\hat{R}$ if R is not fixed. To test the hypothesis of having R change points against the null of no changes, we define the test statistic as

$$S_n = \inf_{\theta}\{MIC(\theta, n, 0)\} - \inf_{\boldsymbol{\theta},\boldsymbol{\tau}}\{MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)\} + Rd \log n, \tag{2}$$

and reject the null hypothesis when S_n is larger than a critical value.

In the next section, we present the result on the limiting distribution of the test statistic S_n under the null hypothesis. We show that S_n diverges to infinity when the alternative model is true. Further, we show that the convergence rate for estimating $\boldsymbol{\tau}$ is $O_p(1)$ and derive the limiting distribution of $\hat{\boldsymbol{\tau}}$. The proofs are presented in Sections 3 and 4, respectively. In the last section, we present some simulation studies.

2. The limiting distribution and convergence rate

Csörgö and Horváth [6] studied the asymptotic distribution of usual likelihood ratio test statistics in single change point case for exponential family. However, the resulting test statistics do not have simple null limiting distributions. In addition, we are not aware of any results in the literature on the null limiting distribution of the usual likelihood ratio test statistic in multiple change point problems. In contrast, we present the simple results on the limiting distribution of S_n in Theorem 1 and the convergence rate and limiting distribution of $\hat{\boldsymbol{\tau}}$ in Theorems 2 and 3, respectively. The proofs will be given in Sections 3 and 4.

Theorem 1. (a) *Under the null hypothesis $H_0 : \theta_1 = \cdots = \theta_R$, Wald conditions $W1$–$W7$ and the regularity conditions $R1$–$R3$, to be specified later in the Appendix, we have, as $n \to \infty$,*

$$S_n \to \chi^2_{(Rd)}$$

in distribution, where d is the dimension of θ and R is the number of change points specified by the alternative hypothesis.

(b) *In addition, if there are R change points at* $\tau_1 = [n\lambda_1], \ldots, \tau_R = [n\lambda_R]$ *with* $0 < \lambda_1 < \cdots < \lambda_R < 1$, *then, as* $n \to \infty$,

$$\inf_{\theta}\{MIC(\theta, n, 0)\} - \inf_{\boldsymbol{\theta},\boldsymbol{\tau}}\{MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)\} \to \infty$$

in probability, which implies that

$$S_n \to \infty$$

in probability.

Theorem 1 implies that the MIC method for testing multiple change points is consistent. That is, when there are R change points in θ at $\tau_1 = [n\lambda_1], \ldots, \tau_R = [n\lambda_R]$ with $0 < \lambda_1 < \cdots < \lambda_R < 1$, the model with R change points will be chosen with probability approaching 1.

Theorem 2. *Under Wald conditions* $W1$–$W7$, *the regularity conditions* $R1$–$R3$ *and the alternative hypothesis* H_1 *that there exist R change points at* $\tau_1 = [n\lambda_1], \ldots, \tau_R = [n\lambda_R]$, *where* $0 < \lambda_1 < \ldots < \lambda_R < 1$, *then we have, for* $r = 1, \ldots, R$,

$$\hat{\tau}_r - \tau_r = O_p(1),$$

where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_R)$ *are defined in* (1) *if R is fixed.*

Obviously Theorem 2 indicates that the estimators $\hat{\tau}_1, \ldots, \hat{\tau}_R$ of the R change points attain the best convergence rate.

Our next theorem is to derive the limiting distribution of the MIC estimator $\hat{\boldsymbol{\tau}}$, which can be characterized by the minimizer of a random walk. Let $\{Y_i^{(r)}, i = \pm 1, \pm 2, \ldots\}_{r=1}^{R}$ be R sequences of independent random variables with $Y_i^{(r)} \sim f(x, \theta_{r0})$ for $i < 0$, and $Y_i^{(r)} \sim f(x, \theta_{(r+1)0})$ for $i > 0$ and $r = 1, \ldots, R$, where $(\theta_{10}, \ldots, \theta_{(R+1)0})$ are the true values of $(\theta_1, \ldots, \theta_{(R+1)})$ under the alternative. For convenience, let $Y_0^{(r)}$ be a non-random number such that $f(Y_0^{(r)}, \theta_{r0}) = f(Y_0^{(r)}, \theta_{(r+1)0})$. Define

$$W_{\mathbf{k}} = \sum_{r=1}^{R}\sum_{j=0}^{k_r} \operatorname{sgn}(k_r)[\log f(Y_j^{(r)}, \theta_{(r+1)0}) - \log f(Y_j^{(r)}, \theta_{r0})]$$

for $k_r = 0, \pm 1, \pm 2, \ldots,$ where $r = 1, \ldots, R$.

With the help of the above notation, the asymptotic distribution of the MIC estimator $\hat{\boldsymbol{\tau}}$ is given as follows.

Theorem 3. *Under the same conditions as Theorem* 2, *we have*

$$\hat{\boldsymbol{\tau}} - \boldsymbol{\tau} \to \boldsymbol{\xi}$$

in distribution, where

$$\boldsymbol{\xi} = \arg \min_{-\infty < k_r < \infty, r=1,\ldots,R}\{W_{\mathbf{k}}\}.$$

The proofs of the theorems will be given in the next two sections.

3. The proof of null limiting distribution

Suppose that the null model is true. That is, all observations in the sequence are independent and identically distributed. In this situation, increasing the model complexity should not boost the maximum possible value of the likelihood function. Our first lemma quantifies this notion. The difference between the maximum values of the likelihood function under the null model and under the alternative model with R change points is no larger than a quantity of order $O_p(\log\log n)$. This result implies that the determining factor for choosing a model is the size of penalty introduced in MIC under the null model. Since the size of penalty is $O(\log n)$, the MIC will select the model with the change points distributed evenly between 1 and n when n increases to infinity.

Lemma 1. *Assume the null hypothesis H_0 is true that there have been no changes in parameters, and the Wald conditions $W1$–$W7$ and the regularity conditions $R1$–$R3$ are satisfied by $f(x,\theta)$. Let θ_0 be the true parameter value of θ. We have*

$$\sup_{\boldsymbol{\theta},\boldsymbol{\tau}} l_n(\boldsymbol{\theta},\boldsymbol{\tau},R) - l_n(\theta_0,n,0) = O_p(\log\log n).$$

Proof. Note that for each given $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$,

$$l_n(\boldsymbol{\theta},\boldsymbol{\tau},R) - l_n(\theta_0,n,0) = \sum_{r=1}^{R+1}\sum_{i=\tau_{r-1}+1}^{\tau_r}[\log f(X_i,\theta_r) - \log f(X_i,\theta_0)].$$

For each non-random τ_r and τ_{r-1},

$$\sup_{\theta_r}\sum_{i=\tau_{r-1}+1}^{\tau_r}[\log f(X_i,\theta_r) - \log f(X_i,\theta_0)]$$

is a usual likelihood ratio statistic. The regularity conditions R1–R3 imply that it converges to a chi-square distribution in distribution when $\tau_r - \tau_{r-1}$ tends to infinity. Hence, each of them is of order $O_p(1)$. Taking maximum over $1 \leqslant \tau_1 < \cdots < \tau_R \leqslant n$ will increase its order to $O_p(\log\log n)$ as shown in [4]. Hence we claim that the lemma is proved. $\square$

When we are forced to fit the data with a model having R change points, the resulting model should still be similar to the null model in some way. In the words of the next lemma, all $R+1$ estimators of θ converge to the true parameter θ_0 under the null hypothesis. This result paves the way for the proof of Theorem 1.

Lemma 2. *Assume that the Wald conditions $W1$–$W7$ are satisfied, the null hypothesis H_0 is true and θ_0 is the true parameter value. Let*

$$\mathcal{S} = \left\{\boldsymbol{\tau} = (\tau_1,\ldots,\tau_R) : \min_{1\leqslant r\leqslant R+1}(\tau_r - \tau_{r-1}) > cn\right\}, \tag{3}$$

where $0 < c < 1$ is a constant. Suppose $\hat{\boldsymbol{\theta}}_{\boldsymbol{\tau}}$ minimizes MIC$(\boldsymbol{\theta},\boldsymbol{\tau},R)$ for given R and $\boldsymbol{\tau}$. Then we have, for each component $\hat{\theta}_r$ of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\tau}}$,

$$\hat{\theta}_r \to \theta_0$$

in probability uniformly for all $r = 1,\ldots,R+1$ and *$\boldsymbol{\tau} \in \mathcal{S}$ as $n \to \infty$.*

Proof. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{R+1}) \neq (\theta_0, \ldots, \theta_0)$, and

$$\mathcal{N}_1 = \mathcal{N}_1(\tilde{\boldsymbol{\theta}}) = \{\boldsymbol{\theta} : (\theta_1 - \tilde{\theta}_1)^2 + \cdots + (\theta_{R+1} - \tilde{\theta}_{R+1})^2 < \rho^2\}.$$

Similar to the proof in [22], we need only show that when ρ is small enough,

$$\max_{\boldsymbol{\tau}\in\mathcal{S}} \sup_{\boldsymbol{\theta}\in\mathcal{N}_1} [l_n(\boldsymbol{\theta}, \boldsymbol{\tau}, R) - l_n(\theta_0, n, 0)] = \max_{\boldsymbol{\tau}\in\mathcal{S}} \sup_{\boldsymbol{\theta}\in\mathcal{N}_1} \sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r} [\log f(X_i, \theta_r) - \log f(X_i, \theta_0)] < 0$$

in probability.

When this is proved, we need only use the compactness of Θ to conclude that $\hat{\theta}_r$ converges to θ_0 in probability. Let, for $\tau_{r-1} < i \leqslant \tau_r$,

$$Y_i^{(r)} = \log f(X_i, \tilde{\theta}_r, \rho) - \log f(X_i, \theta_0),$$

where $f(X, \theta, \rho)$ is defined in condition W2 of Appendix. Since $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_{R+1}) \neq (\theta_0, \ldots, \theta_0)$, there exists at least one r such that $EY_{\tau_r}^{(r)} < 0$ by Jensen's inequality, and all other $EY_{\tau_r}^{(r)} \to 0$ or < 0 when $\rho \to 0$. Assume that $EY_{\tau_{r_0}}^{(r_0)} < 0$ and choose ρ small enough such that all other $|EY_{\tau_r}^{(r)}| < \varepsilon$ for some small $\varepsilon > 0$ (to be specified later). Note that

$$\sup_{\boldsymbol{\theta}\in\mathcal{N}_1} [l_n(\boldsymbol{\theta}, \boldsymbol{\tau}, R) - l_n(\theta_0, n, 0)] \leqslant \sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r} Y_i^{(r)}.$$

Consider the case of $r = 1$. By Kolmogorov maximal inequality [19], that is,

$$P\left\{\max_{1\leqslant k\leqslant n} \left|\sum_{i=1}^{k}(X_i - EX_i)\right| \geqslant \varepsilon\right\} \leqslant \frac{1}{\varepsilon^2}\sum_{k=1}^{n} \operatorname{var}(X_k),$$

if $X_1, \ldots, X_n$ is a sequence of independent random variables with $EX_i^2 < \infty$ for $i = 1, \ldots, n$. Hence,

$$\begin{aligned}\sum_{i=1}^{\tau_1} Y_i^{(1)} &\leqslant \sum_{i=1}^{\tau_1}(Y_i^{(1)} - EY_i^{(1)}) + \tau_1 EY_{\tau_1}^{(1)} \\ &\leqslant \tau_1 \cdot EY_{\tau_1}^{(1)} + o_p(n),\end{aligned}$$

since $E[Y_{\tau_1}^{(1)}]^2 < \infty$ is obvious from condition W2.

Similarly, for $r = 2, \ldots, R+1$, we have

$$\sum_{i=\tau_{r-1}+1}^{\tau_r} Y_i^{(r)} \leqslant (\tau_r - \tau_{r-1}) \cdot EY_{\tau_r}^{(r)} + o_p(n).$$

Hence, we have

$$\begin{aligned}
\max_{\boldsymbol{\tau}\in\mathcal{S}} \sup_{\boldsymbol{\theta}\in\mathcal{N}_1} [l_n(\boldsymbol{\theta}, \boldsymbol{\tau}, R) - l_n(\theta_0, n, 0)] &\leqslant \max_{\boldsymbol{\tau}\in\mathcal{S}} \sum_{r=1}^{R+1} (\tau_r - \tau_{r-1}) E Y_{\tau_r}^{(r)} + o_p(n) \\
&\leqslant (\tau_{r_0} - \tau_{r_0-1}) E Y_{\tau_{r_0}}^{(r_0)} + \varepsilon \max_{\boldsymbol{\tau}\in\mathcal{S}} \sum_{r\neq r_0} (\tau_r - \tau_{r-1}) + o_p(n) \\
&\leqslant \left[c E Y_{\tau_{r_0}}^{(r_0)} + \varepsilon\right] n + o_p(n) \\
&< 0
\end{aligned}$$

in probability, where we choose ε such that $cEY_{\tau_{r_0}}^{(r_0)} + \varepsilon < 0$. Thus the required result follows. □

Remark. In the definition of MIC, we place a penalty term $\sum_{r=1}^{R+1} (\frac{\tau_r - \tau_{r-1}}{n} - \frac{1}{R+1})^2 \log n$ on the likelihood in addition to $(R+1)d \log n$. Lemma 1 implies that MIC is relatively large if $\sum_{r=1}^{R+1} (\frac{\tau_r - \tau_{r-1}}{n} - \frac{1}{R+1})^2$ is larger than some given positive value, as $n \to \infty$. Therefore, the minimum of $MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$ will be reached near $\tau_r = \frac{r}{R+1} n$ for $r = 1, \ldots, R$. Lemmas 1 and 2 together indicate that the MIC value is chiefly determined by the random fluctuation of the likelihood function when $\boldsymbol{\theta}$ is close to its true value and τ_r approximately equals to $\frac{r}{R+1} n$ for $r = 1, \ldots, R$.

We have seen that $\hat{\boldsymbol{\theta}}$ is a consistent estimator of the true parameter $\boldsymbol{\theta}_0$ under the null model when $\boldsymbol{\tau}$ has certain properties. It turns out that the estimator of $\boldsymbol{\tau}$ also has some nice properties.

Lemma 3. *Assume that the Wald conditions $W1$–$W7$ are satisfied. Let $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}})$ be the minimizer of $MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$ for given R. Then under the null hypothesis,*

$$\frac{\hat{\tau}_r}{n} \to \frac{r}{R+1} \quad \textit{for } r = 1, \ldots, R$$

in probability as $n \to \infty$.

Proof. For any $\varepsilon > 0$, define

$$\Delta = \left\{ \boldsymbol{\tau} = (\tau_1, \ldots, \tau_R) : \left| \frac{\tau_r}{n} - \frac{r}{R+1} \right| < \varepsilon, r = 1, \ldots, R \right\}. \tag{4}$$

The lemma is true if we show that $P(\hat{\boldsymbol{\tau}} \in \Delta) \to 1$ when $n \to \infty$. Suppose $\boldsymbol{\theta}_0 = (\theta_0, \ldots, \theta_0)$ and $\boldsymbol{\tau}_R = (\frac{n}{R+1}, \frac{2n}{R+1}, \ldots, \frac{Rn}{R+1})$. Since the penalty term about the locations of change points in MIC disappears if $\boldsymbol{\tau} = \boldsymbol{\tau}_R$ and $l_n(\boldsymbol{\theta}_0, \boldsymbol{\tau}_R, R) = l_n(\theta_0, n, 0)$, it is seen that

$$\begin{aligned}
P(\hat{\boldsymbol{\tau}} \notin \Delta) &\leqslant P\left\{ \min_{\boldsymbol{\tau}\notin\Delta} MIC(\hat{\boldsymbol{\theta}}, \boldsymbol{\tau}, R) \leqslant MIC(\boldsymbol{\theta}_0, \boldsymbol{\tau}_R, R) \right\} \\
&= P\left\{ \max_{\boldsymbol{\tau}\notin\Delta} \left\{ 2l_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\tau}, R) - C \sum_{r=1}^{R+1} \left[\frac{\tau_r - \tau_{r-1}}{n} - \frac{1}{R+1} \right]^2 \log n \right\} \geqslant 2l_n(\boldsymbol{\theta}_0, \boldsymbol{\tau}_R, R) \right\} \\
&\leqslant P\left\{ \max_{\boldsymbol{\tau}\notin\Delta} [l_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\tau}, R) - l_n(\theta_0, n, 0)] \geqslant 4C(R+1)\varepsilon^2 \log n \right\}.
\end{aligned}$$

By the result in Lemma 1,

$$\max_{\boldsymbol{\tau}\notin\Delta}[l_n(\hat{\boldsymbol{\theta}}, \boldsymbol{\tau}, R) - l_n(\theta_0, n, 0)] = O_p(\log\log n).$$

Hence, $P(\hat{\boldsymbol{\tau}} \notin \Delta) \to 0$ as $n \to \infty$. Thus we complete the proof of the lemma. □

With the help of the three lemmas, we are ready to prove Theorem 1.

Proof of Theorem 1. We first prove the theorem for $d = 1$. Lemma 3 tells us that the range of $\frac{\tau_r}{n}$ can be restricted to an arbitrarily small neighborhood of $\frac{r}{R+1}$. When $\frac{\tau_r}{n}$ is restricted to a small neighborhood of $\frac{r}{R+1}$, we have $\boldsymbol{\tau} \in \mathcal{S}$ for some $0 < c < 1$. Thus, we can focus only on $\boldsymbol{\theta}$ in an arbitrarily small neighborhood of $\boldsymbol{\theta}_0 = (\theta_0, \ldots, \theta_0)$ according to Lemma 2.

For any $\varepsilon > 0$ and $\delta > 0$, let Δ be defined as in (4) and define

$$\mathcal{N}_2 = \{\boldsymbol{\theta} : |\theta_r - \theta_0| < \delta, r = 1, \ldots, R+1\}.$$

Let $\hat{\theta}_0$ and $(\hat{\boldsymbol{\theta}}_R, \hat{\boldsymbol{\tau}}_R)$ be the minimizers of $MIC(\theta, n, 0)$ and $MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$ under the restriction $\boldsymbol{\theta} \in \mathcal{N}_2$ and $\boldsymbol{\tau} \in \Delta$. Since the penalty in S_n is always negative, we get

$$S_n \leqslant 2[l_n(\hat{\boldsymbol{\theta}}_R, \hat{\boldsymbol{\tau}}_R, R) - l_n(\hat{\theta}_0, n, 0)] + o_p(1). \tag{5}$$

Our main idea of the proof is to obtain a quadratic expansion for this upper bound in $\hat{\theta} - \theta_0$.

By Taylor expansion at θ_0, we have

$$\begin{aligned}\sum[\log f(X_i, \theta) - \log f(X_i, \theta_0)] = &\sum \frac{\partial \log f(X_i, \theta_0)}{\partial \theta}(\theta - \theta_0) \\ &+ \frac{1}{2}\sum \frac{\partial^2 \log f(X_i, \theta_0)}{\partial \theta^2}(\theta - \theta_0)^2 \\ &+ \frac{1}{6}\sum \frac{\partial^3 \log f(X_i, \xi)}{\partial \theta^3}(\theta - \theta_0)^3\end{aligned} \tag{6}$$

for some $\xi \in \mathcal{N}_2$. The range of summation could be applied to from $i = \tau_{r-1} + 1$ to τ_r or from $i = 1$ to n.

Compared to the quadratic term in (6), the cubic term is negligible when $\delta \to 0$ by condition R2. Let

$$S(X, \theta) = \frac{\partial \log f(X, \theta)}{\partial \theta}$$

be the score function and

$$P_n(\theta, r) = 2\sum_{\tau_{r-1} < i \leqslant \tau_r} S(X_i, \theta_0)(\theta - \theta_0) + \sum_{\tau_{r-1} < i \leqslant \tau_r} \frac{\partial S(X_i, \theta_0)}{\partial \theta}(\theta - \theta_0)^2$$

for $r = 1, \ldots, R+1$. We use $P_n(\theta, 0)$ for the summation from $i = 1$ to n.

By ignoring the cubic term in (6), and using (5) and (6), we get

$$S_n \leqslant \max_{\boldsymbol{\tau}\in\Delta}\sum_{r=1}^{R+1} P_n(\hat{\theta}_r, r) - P_n(\hat{\theta}_0, 0) + o_p(1). \tag{7}$$

This is the quadratic expansion of the upper bound of S_n. We will show that this expansion will lead to a chi-square limiting distribution.

Applying the Kolmogorov maximum inequality [19] again and noting that $\boldsymbol{\tau} \in \Delta$, we have

$$\max_{\boldsymbol{\tau}\in\Delta}\left|\frac{1}{\tau_r-\tau_{r-1}}\sum_{\tau_{r-1}<i\leqslant\tau_r}\frac{\partial S(X_i,\theta_0)}{\partial\theta}+I(\theta_0)\right|=o_p(1), \tag{8}$$

where $I(\theta_0)$ is the Fisher information.

Due to $I(\theta_0) > 0$ and (8), it is obvious that the maximum of $P_n(\theta, r)$ is attained at $\sum_{\tau_{r-1}<i\leqslant\tau_r} S(X_i,\theta_0)/\sum_{\tau_{r-1}<i\leqslant\tau_r}\frac{\partial S(X_i,\theta_0)}{\partial\theta}$ when $n\to\infty$. That is, for $r = 1, \ldots, R + 1$,

$$P_n(\hat{\theta}_r, r)=I^{-1}(\theta_0)\left[(\tau_r-\tau_{r-1})^{-1/2}\sum_{i=\tau_{r-1}+1}^{\tau_r}S(X_i,\theta_0)\right]^2+o_p(1),$$

and

$$P_n(\hat{\theta}_0, 0)=I^{-1}(\theta_0)\left[n^{-1/2}\sum_{i=1}^{n}S(X_i,\theta_0)\right]^2+o_p(1).$$

Without loss of generality, assume that $I(\theta_0) = 1$, and let $Y_i = S(X_i, \theta_0)$ and $W_k = \sum_{i=1}^k Y_i$. Then we have, from (7),

$$\begin{aligned}
S_n &\leqslant \max_{\boldsymbol{\tau}\in\Delta}\sum_{r=1}^{R+1}\left[(\tau_r-\tau_{r-1})^{-1/2}\sum_{i=\tau_{r-1}+1}^{\tau_r}S(X_i,\theta_0)\right]^2-\left[n^{-1/2}\sum_{i=1}^{n}S(X_i,\theta_0)\right]^2+o_p(1)\\
&=\max_{\boldsymbol{\tau}\in\Delta}\sum_{r=1}^{R+1}\left[(\tau_r-\tau_{r-1})^{-1/2}(W_{\tau_r}-W_{\tau_{r-1}})\right]^2-\left[n^{-1/2}W_n\right]^2+o_p(1)\\
&=\max_{\boldsymbol{\tau}\in\Delta}\sum_{r=1}^{R}\left\{\left[\tau_r^{-1/2}W_{\tau_r}\right]^2+\left[(\tau_{r+1}-\tau_r)^{-1/2}(W_{\tau_{r+1}}-W_{\tau_r})\right]^2-\left[\tau_{r+1}^{-1/2}W_{\tau_{r+1}}\right]^2\right\}+o_p(1)\\
&=\max_{\boldsymbol{\tau}\in\Delta}\sum_{r=1}^{R}\left[\tau_{r+1}s_r(1-s_r)\right]^{-1}(W_{\tau_r}-s_rW_{\tau_{r+1}})^2+o_p(1)\\
&\leqslant\max_{\mathbf{t}\in\Delta^*}\sum_{r=1}^{R}T_{nr}^2(t_r)+o_p(1),
\end{aligned} \tag{9}$$

where $s_r = \frac{\tau_r}{\tau_{r+1}}$, $\Delta^* = \{(t_1, \ldots, t_R) : |t_r - \frac{r}{r+1}| < \varepsilon\}$, and

$$\begin{aligned}
T_{nr}(t_r)&=\left\{\frac{[\tau_{r+1}t_r]}{\tau_{r+1}}\left(1-\frac{[\tau_{r+1}t_r]}{\tau_{r+1}}\right)\right\}^{-1/2}\\
&\quad\times\tau_{r+1}^{-1/2}\left\{W_{[\tau_{r+1}t_r]}+(\tau_{r+1}t_r-[\tau_{r+1}t_r])Y_{[\tau_{r+1}t_r]+1}-\frac{[\tau_{r+1}t_r]}{\tau_{r+1}}W_{\tau_{r+1}}\right\}.
\end{aligned}$$

It is obvious that $T_{nr}(t_r)$, $r = 1, \ldots, R$ are asymptotic independent. By Donsker's theorem [7], as $n \to \infty$, for $t_r \in \left[\frac{r}{r+1} - \varepsilon, \frac{r}{r+1} + \varepsilon\right]$, $T_{nr}(t_r) \to [t_r(1-t_r)]^{-1/2} B_{r0}(t_r)$ in distribution as a random continuous function, and $B_{r0}(t)$, $r = 1, \ldots, R$, are R mutually independent Brownian bridges. As a consequence, as $n \to \infty$, we have

$$\sup_{|t_r - \frac{r}{r+1}| \leqslant \varepsilon} T_{nr}^2(t_r) \to \sup_{|t_r - \frac{r}{r+1}| \leqslant \varepsilon} [t_r(1-t_r)]^{-1} B_{r0}^2(t_r)$$

in distribution.

Consequently, from (9) we have shown that

$$S_n \leqslant \sum_{r=1}^{R} \sup_{|t_r - \frac{r}{r+1}| < \varepsilon} T_{nr}^2(t_r) + o_p(1) \to \sum_{r=1}^{R} \sup_{|t_r - \frac{r}{r+1}| < \varepsilon} [t_r(1-t_r)]^{-1} B_{r0}^2(t_r). \tag{10}$$

As $\varepsilon \to 0$, the Lévy modulus of continuity of the Wiener process implies,

$$\sup_{|t_r - \frac{r}{r+1}| \leqslant \varepsilon} \left| B_{r0}(t_r) - B_{r0}\left(\frac{r}{r+1}\right) \right| \to 0$$

almost surely. Since $\varepsilon > 0$ can be chosen arbitrarily small, and

$$\left[\frac{r}{r+1}\left(1 - \frac{r}{r+1}\right)\right]^{-1} B_{r0}^2\left(\frac{r}{r+1}\right) \sim \chi_1^2,$$

(10) implies

$$\varliminf_{n\to\infty} P\{S_n \leqslant x\} \geqslant P\{\chi_R^2 \leqslant x\}$$

for all $x > 0$.

On the other hand, it is straightforward to show that

$$S_n \geqslant \inf_{\theta}\{MIC(\theta, n, 0)\} - \inf_{\boldsymbol{\theta}}\{MIC(\boldsymbol{\theta}, \boldsymbol{\tau}_{\mathbf{R}}, R)\} + Rd \log n$$

$$\to \chi_R^2 \quad \text{as } n \to \infty,$$

where $\boldsymbol{\tau}_{\mathbf{R}} = (\frac{n}{R+1}, \frac{2n}{R+1}, \ldots, \frac{Rn}{R+1})$. Thus,

$$\varlimsup_{n\to\infty} P(S_n \leqslant x) \leqslant P(\chi_R^2 \leqslant x) \quad \text{for all } x > 0.$$

Hence, $S_n \to \chi_R^2$ in distribution as $n \to \infty$.

Consider the case when θ has dimension $d > 1$. The proof for $d = 1$ is also valid up to (6). What we need to pay attention is that Y_k is a vector now. The subsequent order comparison remains the same as the Fisher information is positive definite matrix by the regularity conditions. Therefore, this strategy also works for (9). Then we re-parameterize the model so that the Fisher information is an identity matrix under the null model, and consequently the components of Y_k are uncorrelated. The term $T_{nr}^2(t_r)$ in (9) becomes $T_{nr}^2(t_r, 1) + T_{nr}^2(t_r, 2) + \cdots + T_{nr}^2(t_r, d)$. Also $T_{nr}(t_r, 1), T_{nr}(t_r, 2), \ldots, T_{nr}(t_r, d)$ are asymptotically independent by the central limit theorem for sum of iid random vectors. The remaining proof applies to each of the summands. Hence, we have $S_n \to \chi^2_{Rd}$ in distribution as $n \to \infty$. This proves the conclusion of Theorem 1 under the null hypothesis.

To prove the conclusion of Theorem 1 under the alternative hypothesis H_1. Let $\theta_{10}, \ldots, \theta_{(R+1)0}$ be the true parameter values, not all equal, and $\hat{\theta}$ be the MLE of θ under H_0. Then,

$$
\begin{aligned}
S_n &\geqslant 2\sum_{r=1}^{R+1} \sum_{i=[n\lambda_{r-1}]+1}^{[n\lambda_r]} \log f(X_i, \theta_{r0}) - 2\sum_{i=1}^{n} \log f(X_i, \hat{\theta}) \\
&\quad -C\sum_{r=1}^{R+1}\left(\lambda_r - \lambda_{r-1} - \frac{1}{R+1}\right)^2 \log n \\
&= 2\sum_{r=1}^{R+1} \sum_{i=[n\lambda_{r-1}]+1}^{[n\lambda_r]} [\log f(X_i, \theta_{r0}) - \log f(X_i, \hat{\theta})] + O(\log n).
\end{aligned}
$$

That is, S_n is a sum of $R+1$ likelihood ratio statistics. Each has sample size of order n as it is assumed that $\tau_1 = [n\lambda_1], \ldots, \tau_R = [n\lambda_R]$ for some $0 < \lambda_1 < \cdots < \lambda_R < 1$. Since $\theta_{10}, \ldots, \theta_{(R+1)0}$ are not all equal, $\hat{\theta}$ cannot converge to all of them at the same time. The classical arguments similar to Theorem 1 in [22] implies that

$$
\sum_{[n\lambda_{r-1}]<i\leqslant[n\lambda_r]} [\log f(X_i, \theta_{r0}) - \log f(X_i, \hat{\theta})] \geqslant cn + o_p(n)
$$

for some $c > 0$ in probability for at least one r. For other cases,

$$
\sum_{[n\lambda_{r-1}]<i\leqslant[n\lambda_r]} [\log f(X_i, \theta_{r0}) - \log f(X_i, \hat{\theta})] = O_p(1).
$$

Thus, there exist constants $c > 0$, such that

$$
S_n \geqslant cn + o_p(n) \to \infty,
$$

and also

$$
\inf_{\theta}\{MIC(\theta, n, 0)\} - \inf_{\boldsymbol{\theta},\boldsymbol{\tau}}\{MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)\} = S_n - Rd\log n \to \infty
$$

as $n \to \infty$. Hence we complete the proof of Theorem 1. $\square$

4. The proofs of asymptotic results under alternative

As noticed in the last section, the estimated change points will be forced to distribute evenly between 1 and n under the null model. When the alternative model is true, we might wonder if the MIC estimator of $\boldsymbol{\tau}$ is close to the true value.

In this section, we demonstrate that the MIC estimator of $\boldsymbol{\tau}$ has the best convergence rate (Theorem 2) and derive its limiting distribution (Theorem 3). The key point for proving these results is the consistency of $\hat{\boldsymbol{\theta}}$ upon some conditions. For this purpose, we present that

$\hat{\tau}_r - \tau_r = O_p[n(\log n)^{-1}]$ in the next lemma, where $\tau_1, \ldots, \tau_R$ are the locations of the true change points. These facts further help us to determine the best convergence rate and limiting distribution.

Lemma 4. *Assume that the Wald conditions $W1$–$W7$ and regularity conditions $R1$–$R3$ are satisfied and there exist R change points at $\tau_1 = [n\lambda_1], \ldots, \tau_R = [n\lambda_R]$ with $0 < \lambda_1 < \cdots < \lambda_R < 1$. Then, we have for $r = 1, \ldots, R$,*

$$\hat{\tau}_r - \tau_r = O_p[n(\log n)^{-1}],$$

where $\hat{\boldsymbol{\tau}} = (\hat{\tau}_1, \ldots, \hat{\tau}_R)$ is the MIC estimator satisfying

$$MIC(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\tau}}, R) = \min_{\boldsymbol{\theta},\mathbf{k}} MIC(\boldsymbol{\theta}, \mathbf{k}, R).$$

Proof. For each $r = 1, \ldots, R$, we define

$$A_r(n) = \{\mathbf{k} : 0 < k_1 < \cdots < k_R < n, \text{ and } |k_s - \tau_r| > n(\log n)^{-1}, 1 \leqslant s \leqslant R\}.$$

We claim that $P\{\hat{\boldsymbol{\tau}} \in A_r(n)\} \to 0$, as $n \to \infty$ for $r = 1, \ldots, R$. Since $0 < \lambda_1 < \cdots < \lambda_R < 1$, the claim implies that, with probability approaching 1, exactly one of $\hat{\tau}_1, \ldots, \hat{\tau}_R$ is between $\tau_r - n(\log n)^{-1}$ and $\tau_r + n(\log n)^{-1}$, $r = 1, \ldots, R$. Obviously, this one must be $\hat{\tau}_r$. That is, $\hat{\tau}_r - \tau_r = O_p[n(\log n)^{-1}]$.

To prove the claim, we need only show that

$$P\{MIC(\mathbf{k}, R) > MIC(\boldsymbol{\tau}, R), \text{for all } \mathbf{k} \in A_r(n)\} \to 1.$$

This is true if we show

$$MIC(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{k}, R) - MIC(\boldsymbol{\theta}_0, \boldsymbol{\tau}, R) > Cn(\log n)^{-1} + o_p[n(\log n)^{-1}] \tag{11}$$

uniformly for $\mathbf{k} \in A_r(n)$.

For any $\mathbf{k} = (k_1, \ldots, k_R) \in A_r(n)$, let $\boldsymbol{\theta}^* \in \mathcal{R}^{2(R+1)d}$ be any a vector, and $\mathbf{k}^* \in \mathcal{R}^{2R+1}$ be the vector with the components $k_1, \ldots, k_R, \tau_1, \ldots, \tau_{r-1}, [\tau_r - n(\log n)^{-1}], [\tau_r + n(\log n)^{-1}], \tau_{r+1}, \ldots, \tau_R$, then, by the definition of the maximum likelihood estimator,

$$\begin{aligned} &MIC(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{k}, R) - MIC(\boldsymbol{\theta}_0, \boldsymbol{\tau}, R) \\ &\quad = 2l_n(\boldsymbol{\theta}_0, \boldsymbol{\tau}, R) - 2l_n(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{k}, R) \\ &\qquad + C\sum_{r=1}^{R+1}\left\{\left[\frac{k_r - k_{r-1}}{n} - \frac{1}{R+1}\right]^2 - \left[\frac{\tau_r - \tau_{r-1}}{n} - \frac{1}{R+1}\right]^2\right\}\log n \\ &\quad \geqslant 2l_n(\boldsymbol{\theta}_0, \boldsymbol{\tau}, R) - 2l_n(\hat{\boldsymbol{\theta}^*}, \mathbf{k}^*, 2R+1) + O_p(\log n), \end{aligned} \tag{12}$$

where $\hat{\boldsymbol{\theta}^*} = (\hat{\theta}_1^*, \ldots, \hat{\theta}_{2R+2}^*)$ is the corresponding MLE of $\boldsymbol{\theta}^*$ when there are $2R+2$ segments. Assume that

$$l_n(\hat{\boldsymbol{\theta}^*}, \mathbf{k}^*, 2R+1) = T_1 + \cdots + T_{R+2}, \tag{13}$$

where T_s for $s = 1, \ldots, r-1, r+2, \ldots, R+1$ is the log-likelihood involving $X_i\,(\tau_{s-1} < i \leqslant \tau_s)$, T_r is that involving $X_i\,(\tau_{r-1} < i \leqslant [\tau_r - n(\log n)^{-1}])$, T_{r+1} is that involving $X_i\,([\tau_r + n(\log n)^{-1}] < i \leqslant \tau_{r+1})$, and T_{R+2} is that involving $X_i\,([\tau_r - n(\log n)^{-1}] < i \leqslant [\tau_r + n(\log n)^{-1}])$.

Moreover, let $t(1, s) < \cdots < t(N(s), s)$ denote the elements of the set $\{k_1, \ldots, k_R\} \cap \{\tau_{s-1} + 1, \ldots, \tau_s\}$. Then, for $s = 1, \ldots, r-1, r+2, \ldots, R+1$, by Lemma 1,

$$T_s = \sum_{j=1}^{N(s)+1} \sum_{i=t(j-1,s)+1}^{t(j,s)} \log f(X_i, \hat{\theta}^*_{A(j,s)})$$

$$= \sum_{i=\tau_{s-1}+1}^{\tau_s} \log f(X_i, \theta_{s0}) + O_p(\log\log n), \tag{14}$$

where $t(0, s) = \tau_{s-1}$, $t(N(s)+1, s) = \tau_s$, and $A(j, s) = \sum_{i=1}^{s-1} N(i) + s + j$. Similarly,

$$T_r = \sum_{i=\tau_{r-1}+1}^{[\tau_r - n(\log n)^{-1}]} \log f(X_i, \theta_{r0}) + O_p(\log\log n), \tag{15}$$

$$T_{r+1} = \sum_{i=[\tau_r + n(\log n)^{-1}]+1}^{\tau_{r+1}} \log f(X_i, \theta_{(r+1)0}) + O_p(\log\log n). \tag{16}$$

Also, since $\theta_{r0} \neq \theta_{(r+1)0}$ and $\mathbf{k} \in A_r(n)$ implies that there is no any component of $\mathbf{k}$ between $\tau_r - n(\log n)^{-1}$ and $\tau_r + n(\log n)^{-1}$, by Theorem 1 in [22],

$$T_{R+2} = \max_{\theta} \sum_{i=[\tau_r - n(\log n)^{-1}]+1}^{[\tau_r + n(\log n)^{-1}]} \log f(X_i, \theta) \hat{=} \sum_{i=[\tau_r - n(\log n)^{-1}]+1}^{[\tau_r + n(\log n)^{-1}]} \log f(X_i, \hat{\theta})$$

$$\leqslant \sum_{i=[\tau_r - n(\log n)^{-1}]+1}^{\tau_r} \log f(X_i, \theta_{r0}) + \sum_{i=\tau_r+1}^{[\tau_r + n(\log n)^{-1}]} \log f(X_i, \theta_{(r+1)0})$$

$$- Cn(\log n)^{-1} + o_p[n(\log n)^{-1}]. \tag{17}$$

Hence, by (13)–(17),

$$l_n(\hat{\boldsymbol{\theta}}^*, \mathbf{k}^*, 2R+1) \leqslant l_n(\boldsymbol{\theta}_0, \boldsymbol{\tau}, R) - Cn(\log n)^{-1} + o_p[n(\log n)^{-1}].$$

Thus we get (11) from (12) and hence the claim. This completes the proof. □

Lemma 5. *Assume that the Wald conditions $W1$–$W7$ are satisfied and there exist R change points at $\tau_1 = [n\lambda_1], \ldots, \tau_R = [n\lambda_R]$ with $0 < \lambda_1 < \cdots < \lambda_R < 1$. Assume also that $\hat{\boldsymbol{\theta}}^{(\mathbf{k})}$ minimizes MIC($\boldsymbol{\theta}$, $\mathbf{k}$, R) for each $\mathbf{k}$=$(k_1, \ldots, k_R)$ and given R. Then we have,*

$$\hat{\boldsymbol{\theta}}^{(\mathbf{k})} \to \boldsymbol{\theta}_0$$

in probability uniformly for $|k_r - \tau_r| < n(\log n)^{-1}$ *as* $n \to \infty$, *where* $\boldsymbol{\theta}_0 = (\theta_{10}, \ldots, \theta_{(R+1)0})$ *is the true value of* $\boldsymbol{\theta}$ *under* H_1.

Proof. Define, for $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_R) \neq \boldsymbol{\theta}_0$ and $\rho > 0$,

$$\mathcal{N}_3 = \mathcal{N}_3(\tilde{\boldsymbol{\theta}}) = \{\boldsymbol{\theta} : (\theta_1 - \tilde{\theta}_1)^2 + \cdots + (\theta_{R+1} - \tilde{\theta}_{R+1})^2 < \rho^2\},$$

and

$$\bar{\Delta} = \{\mathbf{k} : |k_r - \tau_r| < n(\log n)^{-1}, r = 1, \ldots, R\}.$$

The lemma is equivalent to that when ρ is small enough,

$$\sup_{\mathbf{k} \in \bar{\Delta}} \sup_{\boldsymbol{\theta} \in \mathcal{N}_3} [l_n(\boldsymbol{\theta}, \mathbf{k}, R) - l_n(\boldsymbol{\theta}_0, \mathbf{k}, R)] < 0 \tag{18}$$

with probability approaching 1.

Note that, for $\mathbf{k} \in \bar{\Delta}$,

$$\begin{aligned} l_n(\boldsymbol{\theta}, \mathbf{k}, R) - l_n(\boldsymbol{\theta}_0, \mathbf{k}, R) &= \sum_{r=1}^{R+1} \sum_{i=k_{r-1}+1}^{k_r} [\log f(X_i, \theta_r) - \log f(X_i, \theta_{r0})] \\ &= \sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r} [\log f(X_i, \theta_r) - \log f(X_i, \theta_{r0})] + o_p(n). \end{aligned}$$

Hence similar to the proof in Lemma 2, we have

$$\sup_{\mathbf{k} \in \bar{\Delta}} \sup_{\boldsymbol{\theta} \in \mathcal{N}_3} [l_n(\boldsymbol{\theta}, \mathbf{k}, R) - l_n(\boldsymbol{\theta}_0, \mathbf{k}, R)] < 0$$

when n is large enough. This completes the proof of the lemma. □

The lemma indicates that we need only focus on a small neighborhood of $\boldsymbol{\theta}_0$ to study the asymptotic properties of MIC when $\mathbf{k}$ is in $\bar{\Delta}$. Now we are ready to prove Theorems 2 and 3.

Proof of Theorem 2. According to Lemma 4, the convergence rate of $\hat{\boldsymbol{\tau}}$ is at least $O_p[n(\log n)^{-1}]$. We now refine the rate based on the initial result.

For any fixed $\varepsilon > 0$, we want to show that there exists $M > 0$, such that

$$P\{|\hat{\tau}_r - \tau_r| > M\} < \varepsilon$$

for n large enough. For this purpose, we define

$$B(n) = \{\mathbf{k} : 0 < k_1 < \cdots < k_R < n, |k_s - \tau_s| < n(\log n)^{-1}, s = 1, \ldots, R\},$$

and

$$B_r(n, M) = \{\mathbf{k} \in B(n) : k_r - \tau_r < -M\}.$$

By Lemma 4, $P\{\hat{\boldsymbol{\tau}} \in B(n)\} > 1 - \frac{\varepsilon}{4}$ for n large enough. Hypothetically, if

$$P\{\hat{\boldsymbol{\tau}} \in B_r(n, M)\} < \frac{\varepsilon}{4}, \tag{19}$$

then for n large enough,

$$P\{\hat{\tau}_r - \tau_r < -M\} \leqslant P\{\hat{\boldsymbol{\tau}} \notin B(n)\} + P\{\hat{\boldsymbol{\tau}} \in B_r(n, M)\} < \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \frac{\varepsilon}{2}.$$

Similarly, $P\{\hat{\tau}_r - \tau_r > M\} < \frac{\varepsilon}{2}$ and hence $P\{|\hat{\tau}_r - \tau_r| > M\} < \varepsilon$.

With the above conclusion, the theorem amounts to show that there exists an M such that (19) holds. For given M and every $\mathbf{k} \in B_r(n, M)$, define

$$\mathbf{l} = (k_1, \ldots, k_{r-1}, \tau_r, k_{r+1}, \ldots, k_R)$$

which belongs to $B(n) - B_r(n, M)$. To prove (19), we need only show that

$$MIC(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{k}, R) - MIC(\hat{\boldsymbol{\theta}}^{(\mathbf{l})}, \mathbf{l}, R) > 0$$

uniformly for $k \in B_r(n, M)$ with probability approaching 1. Note that

$$\begin{aligned} MIC(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{k}, R) &- MIC(\hat{\boldsymbol{\theta}}^{(\mathbf{l})}, \mathbf{l}, R) \\ &= 2[l_n(\hat{\boldsymbol{\theta}}^{(l)}, \mathbf{l}, R) - l_n(\hat{\boldsymbol{\theta}}^{(k)}, \mathbf{k}, R)] \\ &\quad + C\left[\left(\frac{k_{r+1} - k_r}{n} - \frac{1}{R+1}\right)^2 - \left(\frac{k_{r+1} - \tau_r}{n} - \frac{1}{R+1}\right)^2\right] \log n \\ &\quad + C\left[\left(\frac{k_r - k_{r-1}}{n} - \frac{1}{R+1}\right)^2 - \left(\frac{\tau_r - k_{r-1}}{n} - \frac{1}{R+1}\right)^2\right] \log n. \end{aligned}$$

Since $M < \tau_r - k_r < n(\log n)^{-1}$, it is obvious that

$$\left[\left(\frac{k_{r+1} - k_r}{n} - \frac{1}{R+1}\right)^2 - \left(\frac{k_{r+1} - \tau_r}{n} - \frac{1}{R+1}\right)^2\right] \log n = O_p(1)$$

and

$$\left[\left(\frac{k_r - k_{r-1}}{n} - \frac{1}{R+1}\right)^2 - \left(\frac{\tau_r - k_{r-1}}{n} - \frac{1}{R+1}\right)^2\right] \log n = O_p(1).$$

At the same time,

$$\begin{aligned} 2[l_n(\hat{\boldsymbol{\theta}}^{(\mathbf{l})}, \mathbf{l}, R) - l_n(\hat{\boldsymbol{\theta}}^{(\mathbf{k})}, \mathbf{k}, R)] &= 2\sum_{i=k_{r-1}+1}^{k_r} [\log f(X_i, \hat{\theta}_r^{(l)}) - \log f(X_i, \hat{\theta}_r^{(k)})] \\ &\quad + 2\sum_{i=\tau_r+1}^{k_{r+1}} [\log f(X_i, \hat{\theta}_{r+1}^{(l)}) - \log f(X_i, \hat{\theta}_{r+1}^{(k)})] \\ &\quad + 2\sum_{i=k_r+1}^{\tau_r} [\log f(X_i, \hat{\theta}_r^{(l)}) - \log f(X_i, \hat{\theta}_{r+1}^{(k)})] \\ &\hat{=} H_{k1} + H_{k2} + H_{k3}. \end{aligned}$$

By Lemma 5, both $\hat{\theta}_r^{(l)}$ and $\hat{\theta}_r^{(k)}$ converge to θ_{r0}, we may write

$$H_{k1} = 2 \sum_{i=k_{r-1}+1}^{k_r} [\log f(X_i, \hat{\theta}_r^{(l)}) - \log f(X_i, \theta_{r0})]$$
$$-2 \sum_{i=k_{r-1}+1}^{k_r} [\log f(X_i, \hat{\theta}_r^{(k)}) - \log f(X_i, \theta_{r0})],$$

which is the difference between two likelihood ratio statistics. Hence $H_{k1} = O_p(1)$. Similarly, $H_{k2} = O_p(1)$. Now the focus is on H_{k3}, and we write it as

$$H_{k3} = 2 \sum_{i=k_r+1}^{\tau_r} [\log f(X_i, \hat{\theta}_r^{(l)}) - \log f(X_i, \theta_{r0})]$$
$$+2 \sum_{i=k_r+1}^{\tau_r} [\log f(X_i, \theta_{r0}) - \log f(X_i, \hat{\theta}_{r+1}^{(k)})].$$

By Lemma 5, we know that $\hat{\theta}_r^{(l)} \to \theta_{r0}$, $\hat{\theta}_{r+1}^{(k)} \to \theta_{(r+1)0}$. And also note that $\theta_{r0} \neq \theta_{(r+1)0}$, then we choose M large enough such that the second term in the right-hand side of H_{k3} is larger than $CM + M \cdot o_p(1)$ by Theorem 1 in [22], and the first term is $O_p(1)$. That is,

$$H_{k3} \geqslant CM + M \cdot o_p(1).$$

Hence, we have shown that, with probability approaching 1,

$$\min_{\mathbf{k} \in B_r(n,M)} [MIC(\hat{\boldsymbol{\theta}}^{(k)}, \mathbf{k}, R) - MIC(\hat{\boldsymbol{\theta}}^{(l)}, \mathbf{l}, R)] > CM + M \cdot o_p(1) > 0,$$

which implies (19). This completes the proof. □

Proof of Theorem 3. The theorem is equivalent to that, for any given $M > 0$,

$$MIC(\boldsymbol{\tau} + \mathbf{k}) - MIC(\boldsymbol{\tau}) \to 2W_{\mathbf{k}} \tag{20}$$

in probability uniformly for all $\mathbf{k} = (k_1, \ldots, k_R)$ such that $|k_r| \leqslant M$ for $r = 1, \ldots, R$.

Denote $k_0 = k_{R+1} = 0$ for convenience. For all $-M \leqslant k_r \leqslant 0$, we have

$$MIC(\boldsymbol{\tau} + \mathbf{k}) - MIC(\boldsymbol{\tau})$$
$$= 2[l_n(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\tau})}, \boldsymbol{\tau}, R) - l_n(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\tau}+\mathbf{k})}, \boldsymbol{\tau} + \mathbf{k}, R)] + o_p(1), \tag{21}$$

and

$$2[l_n(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\tau})}, \boldsymbol{\tau}, R) - l_n(\hat{\boldsymbol{\theta}}^{(\boldsymbol{\tau}+\mathbf{k})}, \boldsymbol{\tau} + \mathbf{k}, R)]$$
$$= 2 \sum_{r=1}^{R} \sum_{i=\tau_r+k_r+1}^{\tau_r} [\log f(X_i, \hat{\theta}_r^{(\boldsymbol{\tau})}) - \log f(X_i, \hat{\theta}_{(r+1)}^{(\boldsymbol{\tau}+\mathbf{k})})]$$
$$+2 \sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r+k_r} [\log f(X_i, \hat{\theta}_r^{(\boldsymbol{\tau})}) - \log f(X_i, \hat{\theta}_r^{(\boldsymbol{\tau}+\mathbf{k})})]. \tag{22}$$

Since $\hat{\theta}_r^{(\tau)} \to \theta_{r0}$, $\hat{\theta}_{(r+1)}^{(\tau+\mathbf{k})} \to \theta_{(r+1)0}$ by Lemma 5 and $|k_r| \leqslant M$, we have

$$\sum_{r=1}^{R} \sum_{i=\tau_r+k_r+1}^{\tau_r} [\log f(X_i, \hat{\theta}_r^{(\tau)}) - \log f(X_i, \hat{\theta}_{(r+1)}^{(\tau+\mathbf{k})})] = W_{\mathbf{k}} + o_p(1). \tag{23}$$

For the second term in (22), we can easily prove under regularity conditions and Lemma 5,

$$2\sum_{r=1}^{R+1} \sum_{i=\tau_{r-1}+1}^{\tau_r+k_r} [\log f(X_i, \hat{\theta}_r^{(\tau)}) - \log f(X_i, \hat{\theta}_r^{(\tau+\mathbf{k})})] = o_p(1). \tag{24}$$

Hence we get (20) from (21)–(24). The proof is similar when some k_r's are such that $-M \leqslant k_r \leqslant 0$ and others such that $0 \leqslant k_r \leqslant M$. Thus we complete the proof. □

5. Simulation study: the power comparison between MIC and generalized likelihood ratio test

In this section, we conduct a simulation to investigate the finite sample properties of the MIC method applied to two change point problems. We further compare the properties of the MIC and the BIC methods for a couple of penalty constants.

Simulation experiments are done based on four models: normal model with both changes in the mean, normal model with both changes in the variance, exponential model with both changes in the mean, and normal model with both changes in the mean and variance.

The sample sizes of observations are chosen to be $n = 30, 60, 90$, and 120. Under the alternative model we assume that there are two change points in the sequence and place the two change points at $n/6$ and $5n/6$, $n/3$ and $2n/3$, $n/2$, and $3n/4$, and $n/2$, and $2n/3$, respectively. The changes in the normal model are a 0.5 difference in the mean parameter and a factor of 2 in the variance parameter, and in exponential model, the mean parameter change is a factor of $\sqrt{2}$. We choose the nominal levels α as 0.05 and 0.10, respectively. The simulation was repeated 5000 times for each combinations of sample size, location of changes, and so on. To examine the effect of constant C, our simulation was done over a wide range of C including but not limited to $C = 0.0001$, 1, 10, 100, and 1000.

Based on our simulation results, when $C < 1$, both the MIC and BIC methods have very similar power properties. However, the χ^2 distribution is a poor approximation to that of S_n. When $C \geqslant 100$, the χ^2 approximation is good and the power of the MIC is fine, but the estimators of the change points are severely biased toward $n/3$ and $2n/3$ due to the large penalty. Hence we decide to report only the results when $C = 1$ and 10 in the paper. In Tables 1 and 2, we list the powers for both the MIC ($C = 1$ and 10) and BIC methods under the normal levels 0.05 and 0.10, respectively.

Based on the results in Tables 1 and 2, we have the following observations. First, both the MIC and the BIC are consistent, and have higher convergence rates compared to the corresponding methods in the single change point case (see [4]). Second, when the sample size increases, the powers increase significantly for both methods. Third, the MIC method has high powers when $C = 10$ than ones for the method if $C = 1$. Furthermore, there are no any significant differences between the two methods when the two true change points are located at the beginning and the end of the sequence. In other cases, the powers of the MIC are always higher than the powers of BIC for both $C = 1$ or 10. We consider 2% as significant difference with 5000 repetition.

Table 1
Power comparison between MIC and BIC ($\alpha = 0.05$)

τ_1	$n/6$		$n/3$		$n/2$		$n/2$		$n/6$		$n/3$		$n/2$		$n/2$	
τ_2	$5n/6$		$2n/3$		$3n/4$		$2n/3$		$5n/6$		$2n/3$		$3n/4$		$2n/3$	
C	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0
	$n = 30$								$n = 60$							
Normal model: change 0.5 in the mean																
MIC	25.9	25.0	32.2	38.3	22.8	25.1	20.7	23.6	50.9	52.1	58.6	70.1	42.9	51.1	41.9	49.5
BIC	26.0		31.2		22.4		20.3		50.7		55.6		41.2		40.1	
Normal model: change 2 in the variance																
MIC	12.1	12.5	26.3	29.9	32.0	35.6	35.8	40.1	21.9	22.3	51.0	58.5	59.1	64.2	66.9	73.5
BIC	12.2		25.7		31.2		35.3		22.1		50.1		57.7		65.4	
Exponential model: change $\sqrt{2}$ in the mean																
MIC	08.2	08.5	14.9	16.5	15.8	18.2	18.5	21.1	14.2	14.9	30.0	35.2	33.5	37.8	39.4	44.2
BIC	08.3		14.8		15.6		18.4		14.2		28.7		32.3		37.8	
Normal model: changes 0.5 and 2 in the mean and variance																
MIC	13.2	14.4	27.1	32.0	32.5	38.1	37.2	43.5	23.6	23.5	55.4	65.1	64.7	72.4	74.6	81.8
BIC	13.3		26.9		32.5		36.8		23.0		53.9		63.2		73.1	
	$n = 90$								$n = 120$							
Normal model: change 0.5 in the mean																
MIC	72.5	73.2	81.1	88.9	62.8	70.5	59.6	67.8	82.8	84.8	90.1	95.8	75.8	83.8	73.1	80.9
BIC	71.3		78.7		59.8		56.8		82.6		88.5		73.2		70.7	
Normal model: change 2 in the variance																
MIC	33.9	36.1	73.2	81.5	80.0	85.3	85.2	91.0	44.6	46.8	87.0	92.2	89.8	93.3	94.4	96.4
BIC	34.1		71.8		78.3		84.2		45.2		86.1		89.1		93.8	
Exponential model: change $\sqrt{2}$ in the mean																
MIC	17.9	17.4	42.5	47.9	45.3	51.1	54.0	59.9	24.2	24.9	56.1	63.1	60.9	66.4	69.0	75.3
BIC	17.9		41.7		44.3		52.5		24.9		54.8		59.3		67.2	
Normal model: changes 0.5 and 2 in the mean and variance																
MIC	34.7	36.4	79.2	86.8	86.5	91.0	90.9	94.8	49.7	52.8	92.6	96.9	95.3	97.6	98.0	99.3
BIC	34.5		77.6		85.3		89.9		48.4		91.3		94.5		97.6	

Acknowledgments

The authors thank the referees for insightful suggestion on the penalty function and constructive criticism on proofs which lead to a substantially improved paper.

Appendix. Conditions

In this appendix, we present the conditions required in the proof of asymptotic results presented in Sections 2 and 3.

Table 2
Power comparison between MIC and BIC ($\alpha = 0.10$)

τ_1	$n/6$		$n/3$		$n/2$		$n/2$		$n/6$		$n/3$		$n/2$		$n/2$	
τ_2	$5n/6$		$2n/3$		$3n/4$		$2n/3$		$5n/6$		$2n/3$		$3n/4$		$2n/3$	
C	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0	1.0	10.0
	$n = 30$								$n = 60$							
Normal model: change 0.5 in the mean																
MIC	36.9	37.1	44.4	51.3	33.1	36.9	30.3	34.7	63.0	64.1	69.4	79.8	55.3	62.5	53.8	61.5
BIC	36.4		42.2		31.7		28.9		62.4		67.2		53.7		52.1	
Normal model: change 2 in the variance																
MIC	19.7	19.9	36.9	40.8	43.2	46.7	47.3	51.9	32.6	33.7	64.3	71.6	70.1	75.6	77.7	83.0
BIC	19.9		36.2		42.3		46.3		33.1		62.8		69.0		76.3	
Exponential model: change $\sqrt{2}$ in the mean																
MIC	15.3	15.3	25.5	26.5	26.6	29.2	28.6	30.5	23.5	22.6	41.4	45.8	44.2	48.9	50.9	55.9
BIC	15.4		25.1		26.3		28.0		23.3		40.0		42.7		49.0	
Normal model: changes 0.5 and 2 in the mean and variance																
MIC	22.6	23.7	39.6	43.9	46.5	51.7	51.8	57.4	33.9	35.5	68.1	76.2	75.7	81.6	83.2	88.9
BIC	22.7		39.3		46.3		51.2		33.8		66.3		74.0		81.8	
	$n = 90$								$n = 120$							
Normal model: change 0.5 in the mean																
MIC	80.2	81.2	87.1	93.7	72.4	79.5	69.0	77.2	89.7	90.2	94.3	97.8	83.9	90.0	81.7	88.1
BIC	79.1		85.1		70.5		66.6		89.2		93.0		82.2		79.9	
Normal model: change 2 in the variance																
MIC	46.3	47.9	82.4	88.7	87.4	90.7	91.4	94.7	58.4	61.2	92.6	96.1	94.5	96.4	97.2	98.5
BIC	46.2		81.6		86.3		90.4		58.8		91.7		93.5		96.7	
Exponential model: change $\sqrt{2}$ in the mean																
MIC	27.8	28.6	55.2	61.7	58.3	64.7	65.4	71.7	34.7	36.5	68.2	74.4	71.2	77.0	78.4	83.7
BIC	27.6		53.3		56.4		63.4		34.7		66.2		69.5		76.4	
Normal model: changes 0.5 and 2 in the mean and variance																
MIC	47.0	48.8	87.7	93.1	91.9	94.9	94.9	97.1	62.0	65.3	96.0	98.5	97.4	98.9	99.1	99.7
BIC	46.8		86.7		90.9		94.2		61.4		95.5		96.8		98.8	

Suppose $\hat{\boldsymbol{\theta}}_{\boldsymbol{\tau}}$ minimizes $MIC(\boldsymbol{\theta}, \boldsymbol{\tau}, R)$ for given R and $\boldsymbol{\tau}$, then one basic requirement for the solution of change point problems is to estimate the parameters consistently. The MIC is based on the likelihood function, hence it is the minimal requirement to guarantee the consistence of maximum likelihood estimators under iid observations, which is specified in [22]. Consequently, the following conditions look similar to the conditions there.

W1. The distribution of X_1 is either discrete for all θ or is absolutely continuous for all θ.

W2. For sufficiently small ρ and sufficiently large r, the expected values $E[\log f(X, \theta, \rho)]^2 < \infty$ and $E[\log \varphi(X, r)]^2 < \infty$ for all θ, where

$$f(x, \theta, \rho) = \sup_{\|\theta' - \theta\| \leqslant \rho} f(x, \theta') \quad \varphi(x, r) = \sup_{\|\theta' - \theta_0\| > r} f(x, \theta').$$

W3. The density function $f(x, \theta)$ is continuous in θ for every x.

W4. If $\theta_1 \neq \theta_2$, then $F(x, \theta_1) \neq F(x, \theta_2)$ for at least one x, where $F(x, \theta)$ is the cumulative distribution function corresponding to the density function $f(x, \theta)$.

W5. $\lim_{\|\theta\| \to \infty} f(x, \theta) = 0$ for all x.

W6. The parameter space Θ is a closed subset of the d-dimensional Cartesian space.

W7. $f(x, \theta, \rho)$ is a measurable function of x for any fixed θ and ρ.

We will understand the notation E as expectation under the null distribution which has parameter value θ_0 unless otherwise specified.

Furthermore, we require the corresponding regularity conditions [20] since the limiting distribution of S_n is built on the asymptotic normality of the parameter estimators.

R1. For each $\theta \in \Theta$, the derivatives

$$\frac{\partial \log f(x, \theta)}{\partial \theta}, \quad \frac{\partial^2 \log f(x, \theta)}{\partial \theta^2}, \quad \frac{\partial^3 \log f(x, \theta)}{\partial \theta^3}$$

exist for all x.

R2. For each $\theta_0 \in \Theta$, there exist functions $g(x)$ and $H(x)$ (possibly depending on θ_0) such that for θ in a neighborhood $N(\theta_0)$ the relations

$$\left|\frac{\partial f(x, \theta)}{\partial \theta}\right| \leqslant g(x), \quad \left|\frac{\partial^2 f(x, \theta)}{\partial \theta^2}\right| \leqslant g(x), \quad \left|\frac{\partial^2 \log f(x, \theta)}{\partial \theta^2}\right|^2 \leqslant H(x),$$

$$\left|\frac{\partial^3 \log f(x, \theta)}{\partial \theta^3}\right| \leqslant H(x)$$

hold for all x, and

$$\int g(x)\, dx < \infty, \quad E_\theta[H(X)] < \infty \quad \text{for } \theta \in N(\theta_0).$$

R3. For each $\theta \in \Theta$,

$$0 < E_\theta \left\{ \left(\frac{\partial \log f(X, \theta)}{\partial \theta} \right)^2 \right\}, \quad E_\theta \left\{ \left| \frac{\partial \log f(X, \theta)}{\partial \theta} \right|^3 \right\} < \infty.$$

When θ is a vector, the above conditions are assumed true for all components.

References

[1] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, E. Csaki (Eds.), Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973, pp. 267–281.

[2] J. Bai, P. Perron, Estimating and testing linear models with multiple structural changes, Econometrica 66 (1998) 47–78.

[3] D. Barry, J.A. Hartigan, Product partition models for change point problems, Ann. Statist. 20 (1992) 260–279.

[4] J. Chen, A.K. Gupta, J. Pan, Information Criterion and change point problem in regular models, Sankhyā, 2006, accepted for publication.
[5] S. Chernoff, S. Zacks, Estimating the current mean of a normal distribution which is subjected to changes in time, Ann. Math. Statist. 35 (1964) 999–1018.
[6] M. Csörgö, L. Horváth, Limit Theorems in Change-Point Analysis, Wiley, New York, 1997.
[7] M. Csörgö, P. Révész, Strong Approximations in Probability and Statistics, Academic Press, New York, 1981.
[8] Y.-X. Fu, R.N. Curnow, Maximum likelihood estimation of multiple change points, Biometrika 77 (1990) 563–573.
[9] E. Gombay, L. Horváth, An application of U-statistics to change-point analysis, Acta. Sci. Math. 60 (1995) 345–357.
[10] C. Inclán, G.C. Tiao, Use of sums of squares for retrospective detection of changes of variance, J. Amer. Statist. Assoc. 89 (1994) 913–923.
[11] T.-L. Lai, H. Liu, H. Xing, Autoregressive models with piecewise constant volatility and regression parameters, Statist. Sinica 15 (2005) 279–301.
[12] M. Lavielle, Detection of multiple changes in a sequence of dependent variables, Stochastic Process. Appl. 83 (1999) 79–102.
[13] C.-B. Lee, Nonparametric multiple change-point estimators, Statist. Probab. Lett. 27 (1996) 295–304.
[14] C.-B. Lee, Bayesian estimation of the number of change points, Statist. Sinica 8 (1998) 923–939.
[15] Y. Ninomiya, Information criterion for Gaussian change-point model, Statist. Probab. Lett. 72 (2005) 237–247.
[16] E.S. Page, Continuous inspection schemes, Biometrika 41 (1954) 100–116.
[17] E.S. Page, A test for a change in a parameter occurring at an unknown point, Biometrika 42 (1955) 523–526.
[18] G. Schwarz, Estimating the dimension of a model, Ann. Statist. 6 (1978) 461–464.
[19] P.K. Sen, J.M. Singer, Large Sample Methods in Statistics: An Introduction with Applications, Chapman & Hall, New York, 1993.
[20] R.J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.
[21] D. Siegmund, Model selection in irregular problems: applications to mapping quantitative trait loci, Biometrika 91 (2004) 785–800.
[22] A. Wald, Note on the consistency of the maximum likelihood estimate, Ann. Math. Statist. 20 (1949) 595–601.
[23] Y.-C. Yao, Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches, Ann. Statist. 12 (1984) 1434–1447.
[24] Y.-C. Yao, Estimating the number of change-point via Schwarz' criterion, Statist. Probab. Lett. 6 (1988) 181–189.
[25] Y.-C. Yao, S.T. Au, Least-squares estimation of a step function, Sankhyā Ser. A 51 (1989) 370–381.